Towards NLP in Climate Change

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering
by Research

by

Roopal Vaid
201502023
roopal.vaid@research.iiit.ac.in

International Institute of Information Technology
(Deemed to be University)
Hyderabad - 500 032, INDIA
June 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Towards NLP in Climate Change" by *Roopal Vaid*, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Dr. Manish Shrivastava

To my friends and family

Acknowledgments

I'd want to express my gratitude to my adviser, Dr. Manish Shrivastava, for his guidance, encouragement, and support over the years. Professor Radhika Mamidi for helping me in starting my research in ltrc, and giving me an opportunity to explore the field. I am thankful to all of my friends for making this journey memorable. This wouldn't have been possible without their motivation and support throughout the years.

I'm eternally grateful to my parents for their unwavering love and support every step of the way and to Riya, for being a constant source of amusement and inspiration.

Finally, thank you to everyone I met along the way!

Abstract

Climate change is one of the most pressing issues of our time, and understanding the discourse surrounding it is crucial for effective communication and action. The discourse encircling climate change can circumscribe a wide range of perspectives, attitudes and opinions. It is essential to analyze this discourse to identify current challenges, road-maps, and systematic changes governments, organizations, and institutions require to combat the effects of climate change.

Social media is an important platform for climate change discourse due to its widespread use and real-time nature. This makes it possible to analyze the discourse in near real-time, providing valuable insights into public opinions, attitudes, assess topic framing, event dependent attention to the issue and concerns surrounding climate change. We evaluate the contextual and social features that play key role in the coverage on different platforms.

In this thesis, we focus on the fine-grained classification and stance detection of climate change-related social media text surrounding the United Nations Climate Change Conference. We establish two corpora, ClimateStance and ClimateEng with the help of tweets posted during the 2019 United Nations Framework Convention on Climate Change with Intergovernmental Panel in Geneva. We comprehensively outline the dataset collection, pre-processing, annotation methodology, and dataset composition. We have put together a set of guidelines and specifications for creating expandable corpora ClimateEng, ClimateStance which is a collection of 3777 tweets that have been manually labeled with information about events, states, the categories they belong to, and their corresponding stance. We benchmark both datasets for climate change prevention stance detection and fine-grained classification using state-of-the-art methods in text classification and experiments along with results are discussed in detail.

In addition, we create a dataset called ClimateReddit, which is based on Reddit and includes 6262 comments from climate-change related subreddits. We perform semi-supervised learning on the corpus with pseudo-labelling and manually annotate 329 comments for the tasks of fine-grained classification and stance detection of climate-change data. We compare the results with the best-performing models for both tasks from the supervised experiments. Finally, we provide linguistic analysis of ClimateEng, ClimateStance and ClimateReddit using techniques such as part-of-speech tagging and named-entity recognition.

Further, we extend our work in code-mixed setting. We collect Hindi and English code-mixed data from twitter during 2020 and construct a corpus of code-mixed Twitter data. We define the task of fine-

grained classification for the same and outline data-collection and annotation methodology for code-mix data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Climate Change

The consequences of climate change are becoming more noticeable, as natural disasters such as floods, droughts, storms, and fires are becoming more intense and frequent. The changing climate is affecting the biosphere and putting at risk the natural resources and agriculture that are crucial for human survival. According to the IPCC's Fifth Assessment Report [1], continued climate change will have severe and lasting impacts on both people and ecosystems around the world. Climate change and its effects are now a major source of concern globally and have generated increased public discourse. The topic has been widely covered in the news, in scientific journals, and on various social media platforms. Swarnakar and Modi (2021)'s proposes developing a Climate Knowledge Graph which would help in developing a holistic climate policy and functional climate action.

1.2 Social Media Discourse in Climate Change

Assessing public opinion on climate change can be done through various methods, but traditional techniques such as polling do not fully capture the increasing amount of public discourse on the subject in social media. Twitter is a widely used platform and a valuable resource for understanding public opinion and attitudes towards climate change. With over 200 million daily active users and an average annual growth rate of 20%, Twitter provides a wealth of information about public perception on the topic. Linden (2017) discusses the impact social networks have on developing climate change risk perception, which suggests the importance of understanding discourses in social media for this domain.

Discourse on climate change is crucial because it raises awareness of the impact that human activities are having on the climate of the world. People can learn about the science behind climate change, the consequences of inaction, and potential solutions that can help reduce its effects by talking about it. Climate change action will be result of discussion round the issue. Communities will be enabled to

[1] https://www.ipcc.ch/assessment-report/ar5/

understand importance of action and development of programs that promote more sustainable practices and reduce greenhouse gas emissions. Being a global phenomenon, climate change effects each and everyone. Collaboration of different countries, communities, and stakeholders to develop solutions that work for everyone and that can help protect the planet for future generations.

1.2.1 Stance Detection

Discourse stance detection is the extraction of subject's viewpoint, opinion to a claim or a particular topic. The aim is to distinguish whether a given text expresses approval, opposition, or neutrality with respect to a specific subject. Stance detection is important for climate change because it can help to identify and analyze the attitudes and opinions of individuals and groups towards climate change. Stance detection can help to identify individuals or groups who are skeptical or dismissive of climate change, as well as those who are supportive or actively engaged in addressing the issue. This information can be used to target communication and outreach efforts towards different segments of the population, tailoring messages and strategies to be most effective for different groups.

We performed stance detection on English tweets towards the target Climate Change by classifying stance into ternary categories i.e. positive, negative and neutral. Further extended stance classification towards Climate Change in unsupervised setting across Reddit, a different social media platform.

1.2.2 Fine-grained Classification

Fine grained classification is the task of categorizing objects or entities into domain specific subcategories. It provides greater insight into entities and improves the efficiency of classification systems. By providing more challenging and nuanced data-sets, it helps in gaining more detailed information.

One of the most difficult routes at the moment is spreading false information on social media. In online communities, misinformation and echo chambers can be effectively combated with NLP. The spread of misinformation can be intentional as well as unintentional. Algorithms can be trained to detect patterns and false source claims and understand the process of misinformation spread. To limit this we need to further classify our data for information retrieval. Polarization is one of the major issues surrounding the topic of climate change, and As has been observed in previous works, climate change skepticism has achieved a higher level of visibility in media than scientific literature Boykoff and Boykoff (2004). social media plays a significant role in shaping and amplifying these polarized views.

By developing algorithms that can detect and flag false information or identify instances in which people are only being exposed to one perspective, NLP can be used to combat misinformation and echo chambers. NLP can be used to create counter-messaging that presents accurate factual information or uncover diverse perspectives once these instances are identified. Misinformation and echo chambers can have a negative impact on public opinion and tackling this can encourage critical thinking. Cui et al.

(2021) used Köppen classification scheme based on vegetation to project changes in global climate zones.

Since, Climate Change topically covers both effects and contributors of climate change across natural as well as man made phenomenon. We recognize there is a need to further classify our data into Disaster, Forestry, Water, Politics and General categories.

1.3 Code-mixed Text Analysis

Code-mixing is the practice of between two or more languages or language varieties within a single conversation. Particularly observed in multilingual communities, where speakers switch between languages depending on social context and their fluency. It can range from word level to sentence level. It involves expression of cultural identity as well as establishing social relationships.

Sentence: *Ye sab Global Warming ka hi natiza hai ek tarah se agar Manav av environment ko lekar serious nahi hua to aage aur problem face karna parega.*

Words like 'Global Warming, serious' are in English, and words like 'ka', 'natiza', 'parega' etc. are words in Hindi which are transliterated into English.

Sasidhar et al. (2020b) releases Hindi-English code-mixed texts collected from various sources and annotated them with emotions for the task of emotion detection. We extend our work in the domain of Climate Change and released code-mixed Climate Change twitter dataset, with fine grained classification on code mixed data.

1.4 NLP for Social Good

Recently there has been a rise is leveraging NLP systems towards social good. NLP for social good aims to address social issues and contribute towards betterment of lives using machine learning algorithms and language models. NLP is used to build tools towards analysing large chunks of social media text and detect social inclinations, biases. We can track trails of misinformation and keep an eye on public sentiment. Detecting biases can help towards understanding and trace perpetuation of discrimination and power imbalances by representing diverse communities. Advent of tools like high performing machine translators in nlp, text can be translated automatically from one language to the other and improve knowledge sharing across cultures.

Social media platforms such as Twitter and Facebook have given people a powerful tool to express their opinions, engage in discussions, and mobilize others around certain issues. In the case of climate change, these platforms have allowed people to voice their views and opinions, which can range from acceptance of the science behind climate change to skepticism and denial. This diversity of opinions has led to increased polarization on the issue, with people becoming more entrenched in their beliefs and less likely to consider alternative viewpoints.

1.5 Contributions

The focus of this work is centered around stance detection and fine-grained classification of climate change-related social media text. To accomplish this, the key contributions are:-

1. We propose the task of Climate Change prevention stance detection by releasing ClimateStance dataset, an annotated Twitter-based dataset. The dataset consists of 3777 tweets posted during 2019 United Nations Framework Convention on Climate Change.

2. Secondly, we propose the task of fine-grained classification by releasing ClimateEng dataset, a fine-grained classification dataset for climate change related tweets. We also provide the annotation schema for the fine-grained classification.

3. We benchmark both the datasets[2] for climate change prevention stance detection and fine-grained classification using state-of-the-art methods in text classification.

4. We prepare ClimateReddit[3], a Reddit-based pseudo-labeled dataset, from the best performing model for ClimateStance and ClimateEng and benchmark its performance based on a smaller manually annotated test dataset.

5. We perform semi-supervised experiments for both the tasks and benchmark their results using the best-performing model for the supervised experiments.

6. We release a code-mixed dataset on fine-grained classification of climate change related tweets based on three categories: General, Politics, and Awareness.

7. We perform a linguistic feature-based analysis for both ClimateEng and ClimateStance datasets based on part-of-speech tagging and named entity recognition.

1.6 Thesis organisation

The detailed organization of the thesis chapters is as follows:-

1. Chapter 2 presents the related work in the area of Climate Change Discourse Analysis. This provides an outline of various contributions in the domain of climate change social media vox-populi analysis. We discuss state-of-the-art methodologies and experiments.

2. Chapter 3 formally defines the task of stance detection on climate-change related social media text by releasing the ClimateStance dataset. It discusses the brief details of the methodologies and the evaluation metrics used in the bench-marking experiments. We also perform linguistic feature-based analysis for the ClimateStance dataset.

[2] `https://github.com/roopalv54/finegrained-climate-change-social-media.git`
[3] `ClimateReddit`

3. Chapter 4 formally defines the task of fine-grained classification of climate-change related social media text by releasing the ClimateEng dataset based on the five categories: Disaster, Agriculture/Forestry, Ocean/Water, Politics, and General. We also perform linguistic feature-based analysis for the ClimateEng dataset.

4. Chapter 5 describes the semi-supervised experiments performed for the fine-grained classification and stance detection of climate change-related tasks in a Reddit-based dataset: ClimateReddit. We detail our methodology in performing the pseudo-labeling based semi-supervised classification on ClimateReddit. We also release a code-mixed dataset based on three categories: General, Politics, Awareness.

5. Chapter 6 describes the conclusion of the thesis. We discuss our key findings and the potential future work.

Chapter 2

Related Work

With the increased usage of social media there has been significant work towards analyzing climate change, effects of climate change and skepticism towards climate change. Previous work in NLP covers various social media content websites - micro-blogging sites like Twitter, social networking sites like Facebook, Reddit.

2.1 Climate Change in NLP

Early work in analyzing climate-change-related text in the social media setting is primarily focused on statistical analysis Kirilenko and Stepchenkova (2014); Pearce et al. (2013). Kirilenko et al. (2014) collected tweets on climate change and global warming in five languages and studied the effect of geography, time, and major news events that inspired central topics of discussion over climate change. Pearce et al. (2013) presented the tweet authors and topics associated with the publication of the *IPCC's AR5* on the physical science basis for climate change based on the Tweet's hashtags. Moreover, Kirilenko et al. (2014) performed the analysis on tweets during 2012-2013 to conclude that users are establishing a relationship between temperature anomalies and climate change.

On the other hand, Cody et al. (2015) used Hedonometer to determine how collective sentiment differs in response to climate change-related events, news, natural disasters, and oil drilling. They conclude that natural disasters and other phenomena related to climate change contributed to a decrease in overall happiness. Although the works mentioned above are immensely helpful in understanding climate change-related discourses in social media, recent advances in natural language processing enable the fine-grained detection of climate-change-related social media text. The advent of contextualized word representation for improving natural language representation for various downstream tasks, including text classification, has been particularly significant.

Recent work in the area employs the techniques of topic modeling Dahal et al. (2019), and lexicon-based sentiment analysis Loureiro and Alló (2020). Dahal et al. (2019) provided an overview of high-impact areas where machine learning and AI can assist the fight against climate change and highlighted climate mitigation and adaptation, as well as meta-level tools that enable other strategies. Loureiro and

Alló (2020) analyzed Twitter conversations related to climate change in UK and Spain and employed NLP tools to access the sentiment associated and various emotions evoked by these tweets. They used the lexicon developed by the National Research Center Canada (NRC), denoted as EmoLex. used topic modeling to identify topics prevalent related to climate change in a spatio-temporal setting.

2.1.1 Social Media Discourse Analysis

There has been extensive work in the social media discourse analysis in the context of climate change. The prior work by Williams et al. (2015) performs network analysis to identify echo chambers in climate change discourse. Arlt et al. (2018) evaluates the influence of COP21 on climate discourse online. Roxburgh et al. (2019) demonstrates events characteristics and socio-political environment influences climate change coverage on social media platforms.

Roy et al. (2020) gives an insight into social media based communication during a climate disaster like Hurricane Sandy.

2.1.1.1 Stance Detection

With the advancement in the field of stance detection there were contributions to different domain adaptations. Ganin and Lempitsky (2015)'s strategy encourages the development of "deep" features which are invariant towards as well as differentiable for main learning task in the learning domain. They were able to successfully demonstrate the adoption of this method in almost any feed forward network with unlabeled data by augmenting few layers along with a gradient reversal layer. The high rise in the popularity of tweets brought the attention of towards utilizing rich contextual information with linguistic factors to build better twitter sentiment analysis models. They hypothesized the contextual limitations tied to 140 character limit. They explored correlations between these contextual features like geolocation, temporal and author information along with linguistic features and further utilized Bayesian approach to combine this to create a high performing twitter classifier.

Shift in domain results in drop in the best modeled sentiment classifiers. To address this problem of cross domain Pan et al. (2010)'s work marks a bench marking contribution towards cross domain sentiment analysis. They achieve this by spectral feature alignment which aligns words which are domain specific into clusters by employing domain independent words. Domain independent word representations helps them see through domain dependent sentiment analysis.

Barkemeyer et al. (2017) identifies range of contextual factors which co-relate to climate change coverage in newspapers and articles using a cross-sectional regression model. Their findings indicate direct exposure and impact of climate change influences the coverage of climate change in media. Further, there was an increase in coverage seen in areas where there were more regulatory policies and organisations. They successfully highlighted negative influence of unemployment rate and gross domestic product per capita. Moore et al. (2021) provides a systematic review of literature on transformations for climate change mitigation published by addressing the topics of policy relevance and contributes to policy relevant knowledge.

Maynard and Bontcheva (2015) contributed towards an open-source toolkit for enabling researchers to use Twitter to analyze and understand the engagement of the society regarding climate change. Sobhani et al. (2016) released a Twitter dataset for stance detection and further concluded that sentiment features assist in stance classification but are not sufficient on their own.Luo et al. (2020) released the Global Warming Stance Detection Dataset, specifically focused on identifying stance on global-warming-related sentences from news articles. Becerra et al. (2020) performed a geo-spatial analysis of climate change perception through keyword analysis on articles from google scholar and scopus.

2.1.1.2 Fine grained classification

Dasgupta et al. (2018) introduces finer detailed classification of Entity Type. They improve the classification of personal data entities by utilizing the result of existing rule-based annotator systems and use them as complimentary contextual features. Lothritz et al. (2020) performed empirical experiments on transformer-based models vs two non-transformer based models for fine-grained classification of NER in multiple subdomains and showed consistent out-performance of transformer based models in f1 scores and recall values. Suresh and Ong (2021) propose improvement on fine-grained classification tasks by adapting contrastive loss incorporating relationships between labels. Their experiments showed improvement in classification tasks with multiple classes as well as few classes being syntactically close.

2.2 Discourse analysis of Code-mixed Text

Prior work in code-mixed discourse classification, Sasidhar et al. (2020a) incorporated deep neural network to classify emotions in the code-mix dataset. Swami et al. (2018) presents a baseline supervised classification system for stance detection on the code-mix dataset towards demonetization. Das and Gambäck (2014) observes the most erroneous challenges associated with language identification are language boundary detection, reduplication and high code switching points with short format data.

Chapter 3

Fine grained text classification and sentiment analysis

Understanding public opinion and climate change attitudes can be greatly aided by stance detection and text classification. Researchers gain insight into how people perceive and respond to climate change by analyzing large amounts of text data, such as social media posts, news articles, and public opinion surveys.

Researchers can draw inferences about the emotional tone of climate change-related content further on the analysed stance. Policy makers are able to gain insights into the public's attitudes regarding climate change effects and actions towards mitigating climate change by stance detection in various types of text data. These insights can assist public outreach and communication efforts.

3.1 Dataset Creation

This section outlines the dataset creation process for climate change prevention stance detection in climate change related tweets. First, we detail the data collection process, which entails scraping, filtering, and preparing the text for annotation. We then detail the data annotation schema for the climate change prevention stance detection task. Finally, we calculate the inter-annotator agreement to evaluate the efficacy of the annotation process.

3.1.1 Data Collection

Twitter is a social media platform widely used for social and political discussion. This enables real-time user-generated data, making it an important source of information for discourse research. What makes Twitter unique is its focus on user-generated content. User interaction is facilitated, leading to information sharing and lively discussion. User discourse on Twitter can have real-world implications as it is associated with changes in public opinion, political outcomes, and social movements. Studying user discourse on Twitter helps researchers understand how social media influences public opinion and how it can be used to effect change.

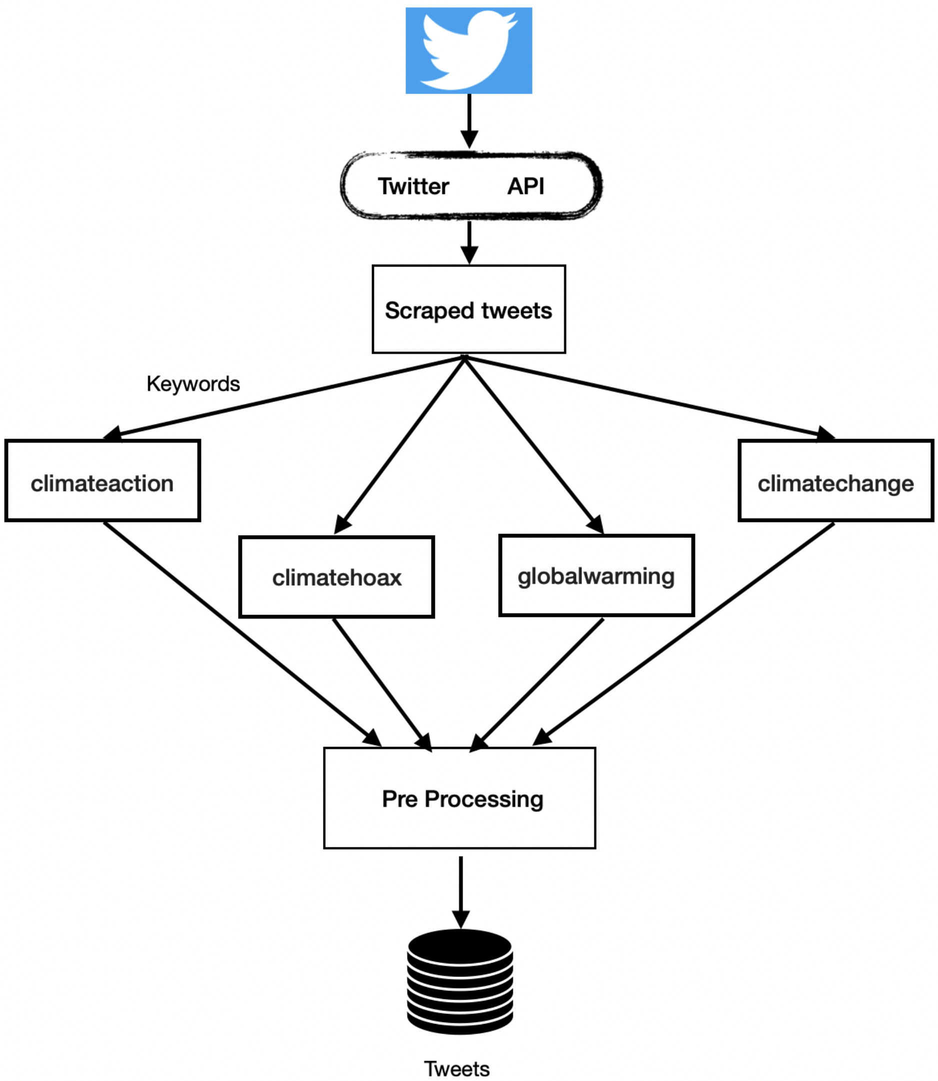

Figure 3.1: Flowchart describing Data Collection Flow used in our Twitter-based Datasets

Climate Change	Global Warming	Warming Planet
#climatechange	#globalwarming	#climatefraud
#fossilfree	#climateaction	#climatetaxfraud
#climatehoax	#climatehoax	Climate Action

Figure 3.2: Twitter keywords used in tweet extraction

We collected a sample of tweets during the UN Climate Change Conference COP 25 held from 2 – 13 December 2019 under the Presidency of the Government of Chile and was held with logistical support from the Government of Spain. The press releases were made available for the press and media advisories through a dedicated webpage. With a wide range of topics covered for climate change technology, innovation, capacity-building, finance, mitigation, and facilitating a platform for local communities and indigenous peoples platform. A live public webcast and coverage was made available. This UN initiative is designed to take the next crucial steps in the next climate change steps. Extending the agreement on Paris Agreement at COP 24 implementation guidelines. The primary goal was to accomplish tasks related to the complete implementation of Paris Agreement. Additionally, the conference helped mobilize the actions of regions, cities, businesses, investors, and civil society by showcasing the wealth of climate action taken by non-Party Stakeholders.

Using the Twitter Application Programming Interface (API) [1], we collected a sample of tweets between 1st December 2019 and 14th December 2019 because the UN Climate Change Conference COP 25 was held from 2 – 13 December 2019, to accommodate different time zones, we start collecting data one day before the conference and collect it until one day after the conference. In total, we collected 378772 tweets along with their metadata. In order to extract climate-change-related tweets from this dataset, we constructed a list of keywords by statistically evaluating the most relevant keywords to the concerns regarding climate change in the extracted twitter dataset from the glossary US EPA* relevant to the concerns regarding climate change - Climate Change, Global Warming, Warming Planet. Apart from these keywords, we also collected tweets containing the associated hashtags #climatechange, #climateaction, #globalwarming, #fossilfree, #climatehoax, #climatetaxfraud.

We primarily focused on the English language. Post removal of non-English tweets, we were left with 263041 tweets. Using twitter ID de-duplication we removed duplicated and overlapping tweets from multiple hashtags and keywords. We removed tweets with offensive language and profanity. Further removed tweets with symbols and emojis. We further de-duplicated the tweets based on urls and tweet text to remove redundant data leaving us with 243781 tweets. Lastly, for performing the human annotation process, we sampled 3777 tweets.

[1] `https://developer.twitter.com/en/docs/twitter-api`

3.1.2 Inter-annotator Agreement

Two human annotators with a linguistic background and proficiency in English conducted the annotation of the dataset to classify the tweets according to the schema mentioned above. We selected a sample annotation set consisting of 100 tweets per class from all across the dataset. Throughout the annotation process, these sample annotation sets served as the reference baseline of each category.

We also analyze the disagreements between the two annotators on the stance detection task. The use of sarcasm in the tweets led to disagreements in many such cases, particularly in the case of stance detection. To accurately capture the stance for those cases, we marked them to be ambiguous. Moreover, the implicit bias of the annotators towards specific entities also led to disagreements between the annotators. We tried our best to select the more objective answer from those labels for creating our corpus.

We calculated the Inter-Annotator Agreement (IAA) to validate the annotation quality. For both annotation tasks, we compute the IAA between the two annotation sets of 3777 tweets using Cohen's Kappa coefficient Fleiss and Cohen (1973). We obtained the Cohen Kappa scores of 0.817 for the *ClimateStance*.

These denote that the quality of the annotations and the presented datasets are significantly productive.

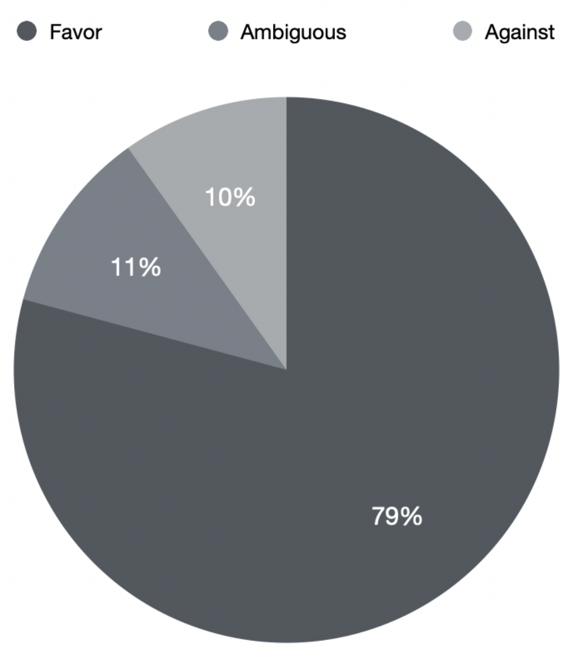

Figure 3.3: Class Composition Chart for ClimateStance Dataset

Class	Example
Ambiguous	*#sustcomm largest journalist collaboration in history to connect news networks covering #ClimateChange* *BBC News - General election 2019: Your questions on climatechange and the environment*
Favor	*I am aware of this, and it does not change the fact that is was a very bad decision regarding CO2 emissions and climate change. The nukes could have acted as backup for quite some time to come, and the coal plants could have been decommissioned a lot earlier.* *It's Independence Day #itsen aisyysp aiv a in #Finland, and we are here on the steps of Parliament demanding #climateaction to keep this country clean and beautifulMillion? Nyt Nyt Nyt#ClimateCrisis #FridaysForFuture #ClimateStrike #NytOnPakko #IlmastoKriisi*
Against	*What's that got to do with climate change? Another one without a straight answer* *Not one climate change sceptic? All we ever hear is a one sided argument for this subject and no one ever has a sceptic / denier. No wonder kids are so brainwashed about this.*

Table 3.1: Examples from the ClimateStance dataset

3.1.3 ClimateStance: Dataset Composition

In the annotated *ClimateStance* dataset, we observe the primary stance to be in *favor* with a count of 2990 (79.16%), i.e., in conclusion, most discussions showed concern and proposed actions to mitigate climate change. Further, we observed the *ambiguous stance* state with no clear stance on climate change 414 (10.96%) times. In contrast, the tweets against and with confusion towards climate change, i.e., those having a *against stance* state, occurred 373 (9.87%) times.

3.2 ClimateStance: Climate Change Prevention Stance Detection

We use the term stance as a broad concept covering evaluation, belief, appraisal, or attitude and its associated information that is stance target and further use this to evaluate the stance. Similar to Sobhani et al. (2016) we use favor, against and ambiguous class labels. We categorize each tweet into one of the three categories in terms of its stance towards climate change prevention:

- **Favor**: Expressions of opinion, action, concern against the climate change phenomenon.
- **Against**: Expressions of distance, ignorance towards signs of climate change, extreme climates, and the opposition of climate change policies or actions taken by the governing bodies.
- **Ambiguous**: Do not express any clear stance towards climate change. Tweets with sarcastic tones were also marked as ambiguous.

Table 3.1 illustrates the example tweets from the ClimateStance dataset.

3.3 Methodology

This section briefly describes the various state-of-the-art models that we used for our bench-marking experiments.

3.3.1 FastText

FastText Joulin et al. (2017) is an open-source, free library from Facebook AI Research (FAIR) for learning word embeddings and word classifications. It allows users to learn text representations and text classifiers and works on standard, generic hardware. Models can later be reduced in size to even fit on mobile devices.

FastText also provides command line arguments for training supervised models with options such as input and output file paths, verbosity level, dictionary options such as minimal number of word occurrences and class occurrences, max length of word ngram, number of buckets, min/max length of char ngram etc.

It allows training both supervised and unsupervised word and sentence representations, also supporting training using both continuous bag-of-words and skip-gram techniques. Since *FastText* uses character n-grams while generating embeddings, it can create representations for words that do not appear in the training corpus. Moreover, *FastText* is capable of achieving good predictive performance efficiently without a pre-trained corpus.

3.3.2 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art pre-training technique for natural language processing (NLP) tasks. The BERT architecture is based on the Transformer model, and it has achieved state-of-the-art results on a wide range of NLP tasks, including question answering, sentiment analysis, and named entity recognition.

Unlike traditional language models, BERT processes text in both directions. It is able to capture more contextual information and hence gains a better language understanding. It is trained with large text corpus sourced from Wikipedia, Google Books etc. Specifically trained for predicting missing words in a sentence by using masked language modeling. Furthermore, it is also trained on a task called next sentence prediction, in which it is trained to predict whether two sentences are related or not.

The BERT architecture primarily consists of two key components: the pre-training stage and the fine-tuning stage.

Pre-training stage: In the pre-training stage, BERT is trained on large amounts of unlabeled text data in a self-supervised manner. The goal of this stage is to learn a general-purpose representation of natural language that can be fine-tuned for specific NLP tasks. The pre-training stage involves two tasks: Masked Language Modeling and Next Sentence Prediction.

In the task of Masked language modeling, BERT takes an input sentence and randomly masks some of the words with [MASK] tokens. The model then tries to predict the masked word's original value based on the surrounding words and the context gained from the other non masked words in the sentence.

For the prediction of the following words requires addition of a classification layer after the encoder output. To transform the output vectors in the vocabulary dimension, the output vectors are multiplied with an embedding matrix. Probability of each word in the vocabulary is calculated with softmax. While calculating the loss function, it only takes into consideration masked values and ignores the prediction of the non-masked words which increases the context awareness.

For Next sentence prediction, BERT takes two input sentences and predicts whether the second sentence is the subsequent sentence in the original text or not. With the assumption of, a random sentence will be disconnected from the first sentence. In the training process half of the inputs consists of pairs where the second sentence is the subsequent sentence and other half the second sentence is randomly chosen from the corpus.

Tokens are used in the beginning of the first sentence, and another token is added to the end of each sentence. Positional embeddings are added to the tokens for the identification of their position in the sequence. The entire input sequence is passed to the transformer and further the probability of IsNextSequence is calculated with softmax.

Fine-tuning stage: In the fine-tuning stage, BERT is fine-tuned on specific NLP tasks using supervised learning. The pre-trained BERT model is used as a starting point, and the model is fine-tuned on a smaller labeled dataset specific to the NLP task.

The BERT architecture consists of a stack of transformer encoder layers. Each encoder layer has two sub-layers: a multi-head self-attention mechanism and a position-wise feed forward network. The multi-head self-attention mechanism allows the model to attend to different parts of the input sequence and generate a context vector for each position in the sequence. The position-wise feedforward network then applies a non-linear transformation to the output of the self-attention mechanism.

In addition, BERT is a bidirectional model, meaning that it processes the input sequence in both directions (from left-to-right and from right-to-left) to generate a more robust representation of the input sequence. Overall, the BERT architecture is a powerful tool for NLP tasks, thanks to its ability to learn high-quality representations of natural language through self-supervised pre-training, and its ability to fine-tune these representations for specific NLP tasks using supervised learning.

Bert-Base and Bert-Large are two popular pre-trained natural language processing models developed by Google. While they share a common foundation, they differ in terms of their size and complexity. Specifically, Bert-Base has 12 transformer layers and 110 million parameters, whereas Bert-Large has 24 transformer layers and 340 million parameters. As a result, Bert-Large is capable of capturing more nuanced patterns in language and can handle more complex tasks, but at the cost of increased computational resources and training time. Therefore, Bert-Base is a more efficient choice for tasks that require a small and fast model, whereas Bert-Large is better suited for more complex language processing tasks.

We use the Base cased and Large cased variants for our benchmarking experiments.

3.3.3 RoBERTa

Roberta and BERT are two different natural language processing (NLP) models that are based on the transformer architecture and use a similar pre-training approach. While they have similarities, there are key differences between the two. Roberta, which stands for ”Robustly Optimized BERT approach”, was designed to capture a wider range of linguistic phenomena by using a much larger training corpus than BERT. Specifically, Roberta was trained on a combined dataset of over 160GB of text, while BERT's training set was 16GB. The larger training corpus enables Roberta to capture more nuances in language and improve its ability to handle a wider range of tasks.

Additionally, Roberta uses a modified training strategy called ”dynamic masking” which encourages the model to learn more generalizable representations of language by preventing it from memorizing specific token patterns. Moreover, Roberta removes the next sentence prediction task from pre-training and increases the maximum sequence length from 512 to 1024. Due to its larger training data and improved training strategy, Roberta often outperforms BERT on a wide range of downstream tasks with the same amount of fine-tuning.

RoBERTa is a transformer-based language model developed by Facebook AI Research (FAIR) that builds upon the architecture of BERT. RoBERTa has two main versions: RoBERTa Base and RoBERTa Large. The main differences between the two versions are in terms of model size, training data, and training duration. RoBERTa Base has 125 million parameters and was trained on 160GB of text data for 4 days, while RoBERTa Large has 355 million parameters and was trained on 800GB of text data for 6 days. RoBERTa Large is more complex and powerful than RoBERTa Base, but it requires more computational resources to use and train. RoBERTa Large is a more powerful and complex language model than RoBERTa Base, with more parameters, larger training data, and longer training duration. However, the increased complexity and computational requirements of RoBERTa Large also make it more challenging to use and train compared to RoBERTa Base.

Unlike *BERT*, it only comes in the cased variant in terms of the type of training data used. We benchmark both Base and large variants of *RoBERTa*.

3.3.4 DistilBERT

DistilBERT Sanh et al. (2019) is a small, fast, cheap and light Transformer model based on the BERT architecture. The size of a BERT model was reduced by 40% via knowledge distillation during the pre-training phase while retaining 97% of its language understanding abilities and being 60% faster. Knowledge distillation is performed during the pre-training phase to reduce the size of a BERT model by 40%. To leverage the inductive biases learned by larger models during pre-training, the authors introduce a triple loss combining language modeling, distillation and cosine-distance losses. Further, it does not use token-type embeddings while removing the pooler in its architecture, reducing the number of layers compared to *BERT* by half. Overall, DistilBERT has about half the total number of parameters of BERT base and retains 95% of BERT's performances on the language understanding benchmark GLUE. In terms of the type of training data used, it can be classified into two variants:- *cased* and *uncased*. We use the cased version of *DistilBERT* for our benchmarking experiments.

Operating the large pretrained models such as BERT and RoBERTa in on-the-edge and/or under constricted computational training or inference budgets remains hard as Transfer Learning from large-scale pre-trained models becomes more prevalent in Natural Language Processing (NLP). With the help of DistilBERT and other distilled models, a smaller general-purpose language representation model can be trained in advance and fine-tuned to perform well on a variety of tasks, much like its larger equivalents.

3.4 Experimental Setting

We evaluate our models on a held-out test dataset for all experiments that consist of 10% of the total dataset. For validation purposes, we split the training dataset was further divided in $8:1$ training:validation split. We use *F1*, *Precision*, *Recall*, and *Accuracy* for evaluating the models. We use the macro variant of the *F1*, *Precision*, and *Recall* which treats all classes equally by taking an unweighted arithmetic mean of all per-class scores.

We use *FastText*'s recently open-sourced automatic hyperparameter optimization functionality and run 100 trials of optimization. For *BERT*, *RoBERTa* and *DistilBERT*, we fine-tune with a learning rate of $1 * 10 - 5$, batch size of 12, and a maximum sequence length of 128 tokens. We validate the models for up to five epochs using the validation dataset and report the best-performing model in our results.

3.4.1 Evaluation metrics

We use the following performance evaluation metrics:-

1. Accuracy is a measure of overall performance of a model.

   $$Accuracy = \frac{TruePositives + TrueNegatives}{TruePositives + FalsePositives + TrueNegatives + FalseNegatives}$$

   It may not give a correct assessment in the case of an imbalanced data.

2. Precision is a measure of how often a model identifies positive instances correctly, in comparison to the total number of identified positive instances.

   $$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

   It is an indicator of false positives and a high precision value is essential in cases where the cost of false positives is high.

3. Recall is a measure of how often a model identifies positive instances correctly, in comparison to total positive instances in the dataset.

   $$Recall = \frac{TruePositives}{(TruePositives + FalseNegatives)}$$

   It is an indicator of identification of positive cases in the dataset and a high recall is essential when the cost of false negatives is high.

4. F1 score is a measure of both precision and recall metrics. To calculate F1 score we take harmonic mean of precision and recall.

   $$F1 = \frac{2 * (precision * recall)}{(precision + recall)}$$

   It ranges from 0 to 1, where the higher score means better precision and recall. It is an essential metric for imbalanced data.

3.5 Results

From Table 3.2 which illustrates the results of the climate change prevention stance detection experiment, we observe *RoBERTa-Base* outperform all models in *F1* with a score of 0.510. In contrast, *RoBERTa-Large* outperforms all models in *Accuracy* and *Recall* with *Accuracy* 82.54% and 0.507 *recall* score. *BERT-LARGE* achieved the best *precision* score of 0.530.

Model \ Metric	F1	Accuracy	Precision	Recall
FastText	0.343	79.63%	0.503	0.354
BERT-Base	0.464	77.51%	0.507	0.446
BERT-Large	0.489	77.78%	**0.530**	0.470
RoBERTa-Base	**0.510**	81.22%	0.528	0.502
RoBERTa-Large	0.489	**82.54%**	0.473	**0.507**
DistilBERT	0.448	79.37%	0.497	0.430

Table 3.2: Results for the Stance Detection using *ClimateStance* dataset

Apart from these, *DistilBERT* and *FastText* also perform competitively while being trained significantly faster than the others. *DistilBERT* obtains an *F1* score of 0.448 in the Climate Change Prevention Stance Detection task. In contrast, *FastText* obtains an *F1* score of 0.343 in the Climate Change Prevention Stance Detection

3.6 Linguistic Feature Analysis

Linguistic feature analysis helps us understand the syntax, morphology, phonology and semantics of languages. It involves examining specific linguistic features in a text corpus. At the syntax level, we can analyze sentence construction and grammatical rules. To understand the corpus's semantics, individual words or phrases must be identified and analyzed in context. Linguistic analysis is crucial for increasing the accuracy of NLP models and is necessary for preprocessing procedures such as tokenization, stopword removal and stemming/lemmatization. It helps locate important textual elements relevant to an NLP task and ensures that the dataset is accurate and free of errors.

We utilize two forms of linguistic feature analysis in our experiments:

1. **Part-of-Speech Analysis**: Part of speech (POS) analysis involves recognizing the grammatical category of words in a sentence based on their syntactic and morphological properties. Various algorithms and tools exist to automatically determine POS tags for words. Pre-trained machine learning models such as HMMs and CRFs are commonly used. POS tagging helps determine the relationship between words and extract contextual information. It can aid in NLP tasks such as parsing, sentiment analysis, machine translation, information extraction and named entity identification. Several POS tagging methods have been established with their own set of tags and criteria. Tagging words with their POS can help comprehend a sentence's grammatical structure, clarify words with multiple meanings and increase the accuracy of NLP tasks.

2. **Named Entity Recognition**: Named Entity Recognition (NER) is a method used to identify and categorize phrases that refer to real-world objects such as people, organizations, locations, dates and times. Algorithms exist to identify and classify these named entities using learning models such as CRFs or RNNs. NER can automatically scan articles and categorize key textual components into predetermined groups. It can be used to extract crucial information for storage in a database or for understanding the text. NER is useful in multiple NLP tasks such as information extraction and question answering. It can also be used in sentiment analysis to associate sentiments with named entities and is critical for machine translation to ensure correct translation of named entities in the target language.

Class	Part-of-Speech					
	PROPN	VERB	NOUN	ADJ	PRON	ADV
Favor	4.22	3.69	7.76	2.00	1.39	1.15
Against	3.22	3.71	7.18	2.26	1.76	1.41
Ambiguous	3.55	3.01	6.47	1.81	1.47	1.16

Table 3.3: ClimateStance Linguistic Feature POS Analysis

We compare our annotated features with various linguistics features including part-of-speech (POS) and named entities (NE) on the *ClimateStance*. To perform this analysis, we exploit SpaCy [2], an open-source library for advanced natural language processing. We use the *en_core_web_sm* for extracting the part-of-the-speech tagging and performing named-entity recognition from all 3777 tweets.

Table 3.3 illustrates the results for the part-of-speech tagging for the *ClimateStance* dataset. We observe that tweets in *favor* stance use proper nouns and nouns the most when compared to other classes. In contrast, tweets with stance *against* displayed a higher use of adjectives, pronouns, and adverbs.

[2] `https://spacy.io/`

The results in Table 3.4 show the mean values of named entity recognition for the *ClimateStance* dataset. While observing NEs, we found the highest occurrence of GPE, MONEY, and ORG tagged NEs in tweets with in *favor* stance. The arguments to support this observation could be stated as in favor stance towards climate change would lead to concern and demand action against climate change. Organizations (ORG) and geopolitical entities (GPE) would be required to make significant changes to bring a systematic change that could slow down climate change. Moreover, the economy needs to adapt to the changing climate, which might be the reason for using entities with the MONEY tag in tweets having stance in *favor* of climate change.

Class	Named Entities				
	PERSON	GPE	MONEY	ORG	DATE
Favor	0.29	0.29	0.39	0.99	0.26
Against	0.30	0.17	0.20	0.71	0.24
Ambiguous	0.31	0.20	0.35	0.82	0.24

Table 3.4: ClimateStance: Mean Value of Named Entities per class

Chapter 4

Towards Fine grained text classification and sentiment analysis of climate change related social media text

News articles, scientific reports, and social media posts are all examples of climate change-related text data that can be categorized using text classification. Researchers may be able to spot patterns or trends in the manner in which the subject is being discussed as a result of this, which can help them gain a deeper comprehension of the distribution of various types of content.

The public discourse regarding climate change can also be monitored and tracked over time with the help of sentiment analysis and text classification. Researchers are able to identify emerging issues and track changes in public attitudes over time by analyzing changes in the types of content produced and the sentiment expressed in that content.

4.1 Dataset Creation

We use the Climate Change data collected from Twitter during the UN Climate Change Conference COP 25 as mentioned in 3.1.1. For Fine grained classification, human annotation same sampled 3777 tweets were used.

4.1.1 Inter Annotator Agreement

Sampling tweets to evaluate inter annotator agreement was performed with the same strategy as discussed in 3.1.2. The Inter Annotator Agreement (IAA) using the Cohen's Kappa coefficient was found to be 0.739 for ClimateEng. For the task of Fine-grained classificaiton on ClimateReddit dataset Cohen Kappa score was calculated to be 0.850.

4.1.2 ClimateEng: Dataset Composition

In the annotated *ClimateEng* dataset, we found the popularity of *General* tweets with a count of 2159 (57.16%) followed by *Politics* class with a count of 1045 (27.67%), which sheds light on how different governing bodies are acting against climate change and citizens' expectations from the governing parties for climate change mitigation. We observed *Ocean/Water* class has a count of 204 (5.40%) as we see the signs of climate change, including rising shorelines and melting glaciers. The *Agriculture/Forestry* class consisted of 197 (5.21%) tweets due to the rising effects of climate change on agricultural practices and biodiversity. We also observed that disastrous events around the globe did follow an increase in discussions regarding climate change and global warming; in the dataset, we were able to capture 172 (4.55%) tweets that could be classified as *Disaster*.

4.2 ClimateEng: Fine-Grained Classification

The collected data was then manually annotated on the following categories: Disaster, Ocean/Water, Agriculture/forestry, Politics, General.

4.2.1 Disaster

This category contains tweets related to various climate-change-influenced natural disasters, including wildfires, floods, hurricanes, and droughts. These references entail:

- References containing opinions about specific instances of natural disasters.
- Information regarding specific instances of natural disasters.

4.2.2 Ocean/Water

This category contains tweets that are:

- References to the effects of climate change on biodiversity on ocean, seas, and other water bodies.
- References to water-based activities that accelerate climate change.
- References to how biodiversity on land adapts to the effects of climate change.

4.2.2.1 Agriculture/forestry

This category contains tweets that are:

- References to the effects of climate change on biodiversity on land, crop yields.
- References to activities including deforestation and fossil fuel burning accelerating climate change.
- References to how biodiversity on land is adapting itself to the effects of climate change.

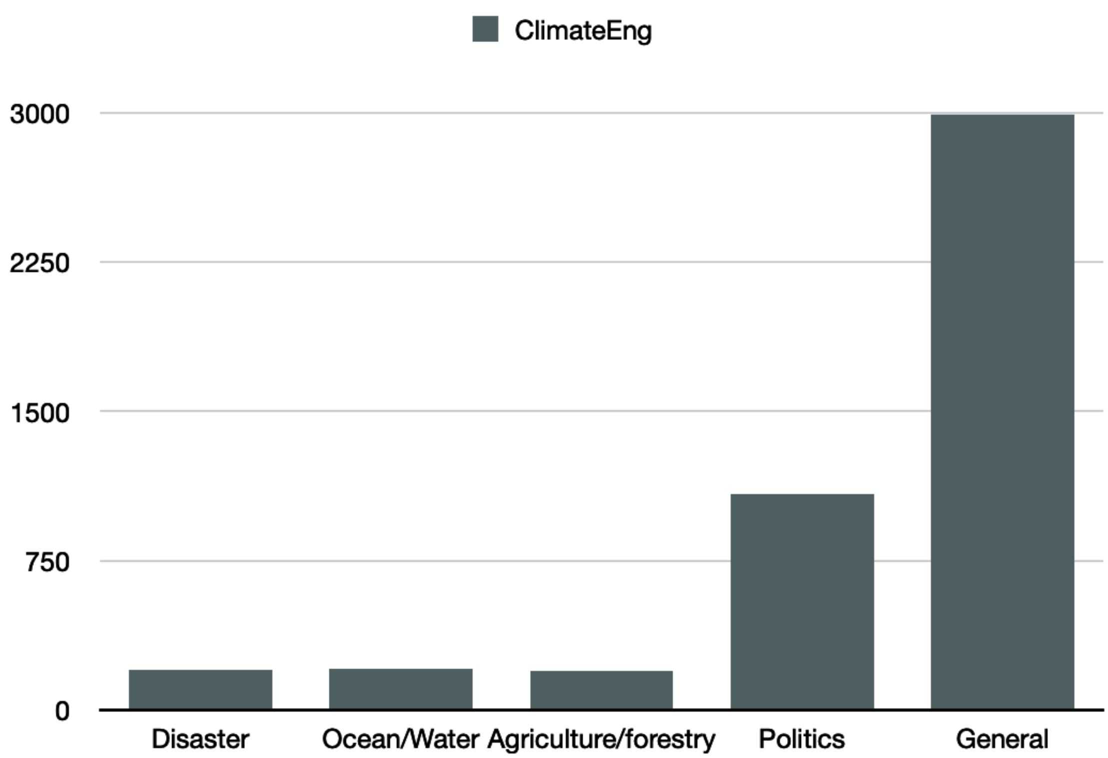

Figure 4.1: Class Composition Histogram for ClimateEng Dataset

4.2.2.2 Politics

This category contains tweets that are related to:

- Quotes of different world leaders on the topic of climate change.
- References about actions taken by institutions like UN to spread awareness about the increasing concerns about climate change.
- References to policies being put in place like Newgreendeal, COP25.

4.2.2.3 General

This category contains tweets that are:

- References of people discussing and spreading awareness about climate change without a specific focus like ocean, water.
- References of climate change affecting suburban lives.

The table 4.1 illustrates the example tweets from ClimateEng dataset

4.3 Methodology

We used the same methodology as mentioned in 3.3. We evaluate the model on the same data split as explained in 3.4. For the experimental settings we used the same values at stated in 3.4.

4.4 Results

The results are evaluated on the Evaluation Metrics defined in 3.4.1

Table 4.2 illustrates the results of the fine-grained-classification experiment. For this task, we observe *RoBERTa-Large* to outperform all models in *F1*, *Accuracy*, and *Precision*, obtaining an *F1* score of 0.735, *accuracy* of 83.07%, and *Precision* of 0.738 in the experiments. At the same time, *RoBERTa-Base* was able to achieve a better *Recall* score of 0.756.

Apart from these, *DistilBERT* and *FastText* also perform competitively while being trained significantly faster than the others.*DistilBERT* obtains an *F1* 0.694 in the fine-grained classification task. In contrast, *FastText* obtains and an *F1* of 0.638 in the fine-grained classification task.

Class	Examples
Disaster	*Take a swim in the charcoal, kids - Sydney beach today (Malabar) #NSWfires #ClimateChange #AustraliaFires* *@AJEnglish Such bunk. IPCC reports have been quite clear, they have little to no confidence that globally, tropical cyclones, floods, drought, tornadoes, SLR, etc, are actually trending worse. And, that's just the data - prior to even considering if man is adding to nothing. #climatechange*
Ocean/Water	*"Scientists have proven that the thinning ice shelves floating around Antarctica are driving ice loss from the interior of the continent as well" - Geophysical Research Letters" #Antarctic #Ice #ClimateChange* *#COP25 – EU Ocean Day highlights the role of oceans in tackling climate change globally*
Agriculture/Forestry	*New report: Two-thirds of North American birds are at increasing risk of extinction from global temperature rise: #climatechange* *Ramon Armengol President elected EU Agricoops #COGECA and from @CoopsAgroES points out the role of livestock farmers and cooperatives to get more added value and fighting the climate change. Cooperatives are key to help farmers to face to current challenges. #Outlook Conference*
Politics	*"It's time for the American electorate to make #climate change a political do-or-die, up and down the ticket." #ClimateCrisis* *Portsmouth Tory candidate misses climate change hustings as she is accused of flying plane across city #GTTO #VoteTactically*
General	*Why do people put 'powered by fairy dust' stickers on their cars? No, you're not, you're powered by fossil fuels that help towards global warming* *"Don't expect the poor to sacrifice for climate change." A scathing look at the climate debate by Swaminathan S Anklesaria Aiyar.*

Model \ Metric	F1	Accuracy	Precision	Recall
FastText	0.638	73.55%	0.730	0.594
BERT-Base	0.696	78.84%	0.697	0.701
BERT-Large	0.695	78.31%	0.730	0.689
RoBERTa-Base	0.734	80.16%	0.725	**0.756**
RoBERTa-Large	**0.735**	**83.07%**	**0.738**	0.742
DistilBERT	0.694	77.51%	0.695	0.713

Table 4.2: Results for the Fine-grained classification using *ClimateEng dataset*

4.5 Linguistic Feature Analysis

We perform a similar methodology for linguistic analysis for ClimateEng as we did in the Section 3.6.

Table 4.3 illustrates the results for the part-of-speech tagging for the *ClimateEng* dataset.

Table 4.4 illustrates the results for named entity recognition for the *ClimateEng* dataset. Tweets classified as *Disaster* had the majority of GPE NEs as well as DATE NEs. We believe this could be due to the localization of disastrous events and tweets holding the political body of the geography for action for mitigation and relief work. Tweets classified as *General* observed the least mention of MONEY NEs. In contrast, we see a higher count of the MONEY NEs in *Agriculture* and *Disaster* classes, which might be due to the cost associated with agricultural industries and disaster mitigation and relief organizations to adapt to the climate change effects witnessed during a disaster. We also observe the most leading mention of ORG NEs in *Politics* class. This observation could be due to references of actions needed to be adopted or are adopted by different organizations to mitigate climate change.

Class	Part-of-Speech					
	PROPN	VERB	NOUN	ADJ	PRON	ADV
General	3.58	3.33	7.08	1.92	1.47	1.13
Politics	4.75	4.26	8.14	2.21	1.68	1.34
Ocean/Water	4.94	3.17	7.96	1.78	0.82	0.86
Agriculture/Forestry	4.10	3.51	8.73	1.92	0.72	0.95
Disaster	4.46	3.81	8.24	2.23	1.09	1.34

Table 4.3: ClimateEng Linguistic Feature POS Analysis

Class	Named Entities				
	PERSON	GPE	MONEY	ORG	DATE
General	0.29	0.16	0.35	0.85	0.23
Politics	0.38	0.41	0.39	1.12	0.28
Ocean/Water	0.22	0.40	0.37	0.97	0.32
Agriculture/Forestry	0.16	0.22	0.47	0.95	0.19
Disaster	0.27	0.61	0.42	0.95	0.35

Table 4.4: ClimateEng: Mean Value of Named Entities per class

Figure 4.2: Word Cloud for ClimateEng and ClimateStance Dataset

Chapter 5

Semi-Supervised Learning

5.1 Semi supervised learning

Semi-supervised learning is a machine learning approach that improves model performance by combining labelled and unlabeled data. Semi-supervised learning in natural language processing (NLP) may be used to train models on vast volumes of unlabeled text data and then fine-tune the model using a smaller quantity of labelled data.

A model is trained on a labelled dataset with each sample linked with a label or category in classical supervised learning. Based on the patterns acquired from the labelled data, the model learns to categorise new cases. Labeled data, on the other hand, can be costly and time-consuming to obtain, particularly in NLP, where human labelling of text data can be difficult. In contrast, semi-supervised learning allows models to be trained on both labelled and unlabeled data. The model can learn patterns and features from unlabeled data that may be used for classification tasks. The model may then be fine-tuned with less labelled data to change its predictions to meet the precise categories or labels of interest.

5.2 Self-training with Pseudo-Labelling

There are various semi-supervised learning strategies, including Self-training, where a model is first trained on a smaller labelled dataset. The labelled dataset is first split into train and test sets for the same. The larger unlabeled data is then classified using the best performing trained classifier model, and the instances with the highest confidence ratings are concatenated with the labelled dataset. After that, the model is retrained on the larger labelled dataset, and the procedure is continued until convergence is reached.

It enables models to be trained on greater volumes of data, even unlabeled data, which improves the model's efficiency and accuracy. Allows models to learn from unlabeled data, lowering the cost of data labelling. Furthermore, it is more resilient and able to generalize to new data than models trained just on labelled data. As a result, semi-supervised learning is an effective approach for training NLP models on vast volumes of data, including unlabeled data, and it can result in better performance and efficiency.

5.3 ClimateReddit

5.3.1 Semi supervised dataset collection

Reddit is a valuable data source for analyzing social discourse, especially in the study of online communities, politics and social issues. Firstly, reddit is one of the most popular social media platforms with a diverse user base from different age groups, socio economic and geographic backgrounds. Secondly, subreddits provide community groups focused on specific topics. Futher reddit users leverage anonymity which helps them accurately express their thoughts and opinions within those communities. Reddit is known for its lengthy content and detailed user posts and comments on fine details of topics addressed within these subreddits. Reddit provides a real time platform where users can participate in real time discussions and debates which provides valuable data to study social discourse.

For the task of extraction of Reddit comments related to climate change we use Pushshift by Baumgartner et al. (2020). For this purpose, we use four primarily active subreddits that engage in climate change discourse, namely: *r/climate*, *r/Climateskeptics*, *r/ClimateActionPlan*, *r/climatechange*. Through this method, we extracted 6591 climate change related user comments in total. We then preprocessed these comments to remove hyperlinks and markdown symbols representing stylized text (i.e., bold and italic). Finally, we split the dataset into two parts: 6262 comments for creating the pseudo-labeled dataset and 329 comments for manual annotation for bench-marking the pseudo-labeled dataset.

5.3.2 Methodology

We create *ClimateReddit* dataset to perform experiments with semi-supervised learning for the task of Stance Detection and Fine-grained Classification for the Reddit-based dataset. Semi-supervised learning is often used for utilizing a large amount of unlabeled data to improve the predictive performance of models across various machine learning tasks Blum and Mitchell (2000); Chapelle et al. (2006). For our semi-supervised experiments, we use the method of pseudo-labeling. In this method, we first train a "teacher" model based on our Twitter-based annotated datasets, namely, *ClimateEng* and *ClimateStance*. We then use this model to predict the labels for the un-annotated Reddit dataset. This results in a pseudo-labeled dataset from the predictions made by the best performing models on the ClimateEng and ClimateStance datasets. We denote this pseudo-labeled dataset of Reddit comments along with its predicted Stance and Fine-grained climate-based classification labels as *ClimateReddit*.

The workflow for semisupervised experiments with pseudo labeling is presented at 5.1

5.4 Experiments and Results

For generating pseudo-labels and performing the bench-marking experiments, we use the best-performing model in terms of *F1* score for both tasks of stance detection and fine-grained classification of climate change text. We use a 10% of the total dataset as our test data. By spliting the training dataset

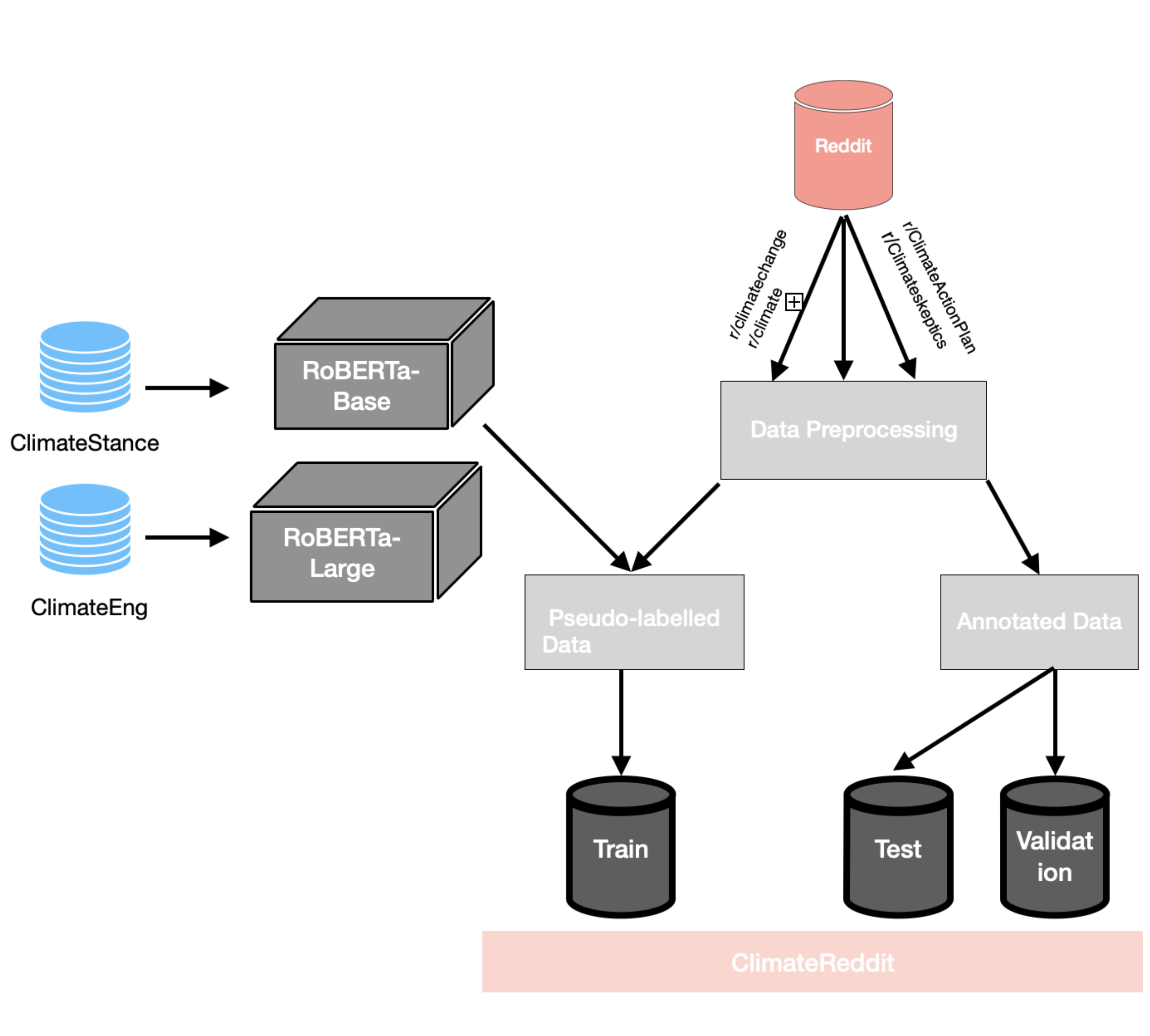

Figure 5.1: Workflow of semi-supervised learning with ClimateReddit

into a ratio of $8 : 1$ we generate a training:validation split. Our evaluation metrics include *F1 score*, *Precision*, *Recall*, and *Accuracy*.

We use all splits of the Twitter-based datasets, namely *ClimateStance* and *ClimateEng*, for their respective tasks, for training the generating the pseudo-labels from the Reddit dataset. For validation, we re-split the dataset into a $9 : 1$ split. Now, upon pseudo-labeling, we use the aggregated dataset consisting of both Twitter and Reddit text and re-split the dataset again into a $9 : 1$ split for validation. For all our evaluation experiments, we use the same manually annotated dataset split of *ClimateReddit* as the test dataset.

For this experiment, we use the best performing models in terms of *F1* score for the Climate Change Prevention Stance Detection task using *ClimateStance* (*RoBERTa-Base*) and Fine-grained Classification task using *ClimateEng* dataset (*RoBERTa-Large*).

Metric / **Training Data**	**F1**	**Accuracy**	**Precision**	**Recall**
ClimateEng	0.775	88.15%	0.800	0.769
ClimateEng + *ClimateReddit*	0.834	90.27%	0.850	0.823

Table 5.1: Results for the Semi-Supervised Fine-grained Classification task

From Table 5.1, for the task of fine-grained classification, we find that *RoBERTa-Large* trained with all splits of *ClimateEng* performs significantly well in the fine-grained classification task for ClimateReddit dataset, obtaining an F1 of 0.775 and an accuracy of 88.15%. Moreover, using the pseudo-labeled Reddit dataset for training along with *ClimateEng*, we find an even higher F1 of 0.834 and an accuracy of 90.27%.

Metric / **Training Data**	**F1**	**Accuracy**	**Precision**	**Recall**
ClimateStance	0.343	60.79%	0.403	0.387
ClimateEng + *ClimateReddit*	0.311	60.49%	0.396	0.369

Table 5.2: Results for the Semi-Supervised Stance Detection task

In contrast, as illustrated in Table 5.2, the predictive performance of *RoBERTa-Base* reduces sharply for the task of Stance detection in the semi-supervised setting. It obtains an F1 score of 0.343 and an accuracy of 60.7% when only trained with the *ClimateStance* dataset. Upon adding the additional Reddit-based pseudo-labeled corpus for the Stance detection, we find the model's performance to dip even further, reaching an F1 score of 0.311 and an accuracy of 60.49%. This drop can be attributed to the significant imbalance in class distribution.

In the annotated *ClimateEng* dataset, we found the popularity of *General* tweets with a count of 2159 (57.16%) followed by *Politics* class with a count of 1045 (27.67%), which sheds light on how different governing bodies are acting against climate change and citizens' expectations from the governing parties for climate change mitigation. We observed *Ocean/Water* class has a count of 204 (5.40%) as we see the signs of climate change, including rising shorelines and melting glaciers. The *Agriculture/Forestry* class consisted of 197 (5.21%) tweets due to the rising effects of climate change on agricultural practices and biodiversity. We also observed that disastrous events around the globe did follow an increase in discussions regarding climate change and global warming; in the dataset, we were able to capture 172 (4.55%) tweets that could be classified as *Disaster*.

Figure 5.2: Word Cloud for ClimateReddit dataset

In the *ClimateReddit* dataset consisting of 6591 Reddit comments, we observe the primary stance to be in favor with a count of 6269 (95.11%). Further, we observed the *against* stance 251 (3.80%) times and those having a *ambiguous* stance, occurred 71 (1.08%) times. Moreover, upon observing in terms of the fine-grained labels, we found 4699 (71.29%) comments to lie in the *General* category. The next most frequent category was *Politics*, having 1197 (18.16%) comments. The next three categories of comments had a fairly equivalent number of occurrences having 243 (3.69%), 227 (3.44%), and 225 (3.41%) comments for *Ocean/Water*, *Agriculture/Forestry*, and *Disaster* respectively.

5.5 Code-mix Dataset and Experiments

A code-mixed dataset is a set of language samples where speakers use more than one language in a single utterance. This is a common phenomenon in multilingual societies, where people mix elements from different languages in their speech. Code-mixed data is particularly used in natural language processing and machine learning applications, such as language identification, machine translation, and sentiment analysis. NLP models trained on monolingual datasets may not perform well on code-mixed data. By creating and using code-mixed datasets, we can improve the performance of their models on real-world data, especially for applications that target multilingual audiences.

Ahmad and Singla (2021)'s work in sentiment analysis of code-mixed social media data highlights the challenges of working of working with code-mixed data in comparison to monolingual data. The findings concluded noisy nature of corpus due to spelling variations, slangs, abbrevations which challenges the current best performing models trained on monolingual data. Common errors were enhanced due to mis-identification of language at word level.

5.6 Code-mix Dataset Annotation and Data Processing

We extracted code mixed tweets from twitter. We used twitter API for tweet extraction. The data collection was performed for the year of 2021 and extracted twitter dataset from the glossary of hashtags #climatechange, #climateaction, #globalwarming, #fossilfree, #climatehoax, #climatetax to identify tweets related to climate change.

We perform token level language detection using spacy. For each tweet, we classify it as codemix if the number of tokens classified as Hindi and number of tokens classified as English both exceed 0.2 fraction in the tweet. After identification of these code mixed tweets, we performed preprocessing which included removing profanity, urls and deduplication with tweet ids. Futher we annotated this data with two annotators. Both of the annotators speak and write Hindi and English. The annotation task was performed for fine-grained classification.

5.6.0.1 Fine-grained Classification

For the task of Fine-grained classification the following classes were identified:-

- **Politics**
  This class of tweets were found to be addressing and raising concerns regarding public policies and political discourse regarding climate change.

  Examples:-

  1. *Government se nivedan hai ke global warming ke lie na sahi inke lie Trees laga de*
  2. *PM Modi:- Global warming nahi hai mitrooon... ey toh clouds hai jo hamare soch ko blur kar rahe hai.*

- **Awareness**

  The category of tweets were identified spreading awareness and information regarding climate change and global warming.

  1. *Aj ka DNA bahut acha tha Mausam ki jankari Global warming ka jo details aap ne di Bahut acha laga*
  2. *Ye sab Global Warming ka hi natiza hai ek tarah se agar Manav av environment ko lekar serious nahi hua to aage aur problem face karna parega*

- **General**

  The rest of the tweets were classified into general category. They were found to be addressing climate change, global warming in a generic context.

  1. *nahi , Global warming ka kasoor hae..*
  2. *Hum the jinke sahare , Kat gaye vo ped saare !! #GlobalWarming #GoGreen #savetree*

5.6.0.2 Dataset composition

In the annotated code mix dataset, the most frequent occurrence was of the class *General*. The General class composed of 1115 tweets i.e. 68.07% of the dataset while Politics composed of 16.12% with the count of 264 and Awareness composed of 259 tweets taking up 15.81%.

Chapter 6

Future Work and Conclusions

6.1 Conclusion

In this work, we proposed the task of predicting Stance in social media texts related to climate change. We benchmarked the datasets using state-of-the-art contextualized word embeddings and provided baselines for the proposed task. Following observation found that *RoBERTa-Base* obtained the best *F1* score in the Stance detection task with a 0.510 *F1* score. We outlined each stage of the pipeline, including dataset creation, and experminent modeling.

We further introduce the task of Fine-grained classification of social media text into the following five categories: *Disaster*, *Forestry*, *Ocean*, *Politics* and *General*. We provide baseline for the task by using the contexualized word embeddings and bench mark the ClimateEng dataset. Moreover, we observed that *RoBERTa-Large* outperforms all other models in three of the four evaluation metrics for the fine-grained classification task, obtaining an *F1* of 0.735.

We further extended this work to the semi-supervised setting. For this purpose, we created the dataset ClimateReddit composed of comments from climate related subredits. We use the best performing Teacher models from ClimateStance i.e. *RoBERTa-Base* and ClimateEng i.e *RoBERTa-Large* for pseudo-label generation of ClimateReddit for Stance detection and Fine-grained classification tasks. We saw improved F1 score as 0.83 and accuracy as 90.27% on fine grained classification task.

We further examined linguistic features such as POS and NER on both the datasets using SpaCy with the help of model's NLP characteristics. From our results, we observed in favor tweets had the highest usage of nouns and proper nouns. The most frequent occurrences of NEs was found in favour stance. Text in favor stance is more accountable towards climate change action, which leads to higher occurrence of GPE, MONEY and ORG NE's. Due to geo political localization of disastrous events, the disaster class had the maximum occurrences of NE's. Verb tags were consolidated highly in politics class, and thus Climate action is identified in Politics class.

This analysis of linguistic features can be further extended to entail a study on the correlation of these features alongside fine-grained labels and stance labels created in *ClimateStance* and *ClimateEng* dataset. The study may lead to interesting sociolinguistic findings while helping out in general understanding of how we use language in a social setting while writing climate-related short-form text. Moreover, this study may also help with information retrieval Li et al. (2022) based on the named entities alongside our created labels.

We further extended our analysis on code-mixed dataset. This dataset consists of code-mixed tweets identified towards the target of Climate Change. The fine grained classification resulted in *Politics*, *Awareness* and *General* classes.

6.2 Challenges

One of the major challenges we faced with social media discourse analysis is decreased time value of social media data. Since the data across different social media platforms is generated at a high volume and high velocity, it makes it challenging to store and analyze data in real time. We also noted that with the case of statistical models, domain knowledge is required to annotate training data which is time expensive. This requirement substantially increases in the case of climate change domain to prevent the model performing poorly due to the absence of contextual domain knowledge associated with tokens.

6.3 Future work

As social media platforms have grown so rapidly in the last ten years, a lot of unstructured text data has been generated. Researchers now have a wealth of opportunity to examine and comprehend human behaviour, social relationships, and attitudes because to the availability of this enormous amount of data. Applications like opinion mining, sentiment analysis, and content suggestion all heavily rely on the fine-grained categorization of social media data.

In conventional methods of text categorization, domain-specific characteristics are manually extracted in order to represent the text data. These methods, however, take a lot of effort and demand a high level of topic knowledge. With the introduction of deep learning, the emphasis has turned to creating fully trainable models that can detect the pertinent elements from the data automatically. As an extension, we can develop models in **Domain Agnostic** settings

As NLP models are taught to recognise the patterns and structure of language, linguistic analysis of the dataset is a crucial step towards increasing accuracy. It is simpler for NLP models to correctly learn these patterns and structures when linguistic analysis is used to assist detect them. It is crucial to preprocess the text by carrying out operations such as tokenization, stopword removal, stemming/lemmatization, before applying any NLP algorithm to a dataset. Linguistic analysis is necessary for these preprocessing procedures in order to recognise and comprehend the language's laws and patterns. It aids in locating crucial textual elements that are pertinent to a certain NLP assignment. Moreover, it makes sure that the dataset is accurate and free of faults like grammatical or spelling issues.

This work can be further expanded to analyze social reaction towards climate change in multilingual settings to get a broader understanding of discourse in Climate Change across different language groups.

We can further experiment with **Teacher-Student multi-lingual model learning** for low resource language.

We can extend this work to perform **fine-grained stance analysis** on the Climate Change datasets. Schuff et al. have contributed towards establishing correlation between fine-grained emotions on a stance detection corpus. By analysing sentence by parts, we can apply stance analysis on a sub-sentence level. Identification of topic i.e. target of the sentiment and the sentiment will give a more nuanced stance analysis and what linguistic and social characteristcs are deriving those stance values.

Related Publications

- **Roopal Vaid**, Kartikey Pant, and Manish Shrivastava. 2022. Towards Fine-grained Classification of Climate Change related Social Media Text. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 434–443, Dublin, Ireland. Association for Computational Linguistics

Bibliography

Gazi Imtiyaz Ahmad and Jimmy Singla. Sentiment analysis of code-mixed social media text (sa-cmsmt) in indian-languages. In *2021 International Conference on Computing Sciences (ICCS)*, pages 25–33, 2021. doi: 10.1109/ICCS54944.2021.00014.

Dorothee Arlt, Imke Hoppe, Josephine B. Schmitt, Fenja De Silva-Schmidt, and Michael Brüggemann. Climate engagement in a digital age: Exploring the drivers of participation in climate discourse online in the context of cop21. *Environmental Communication*, 12(1):84–98, 2018. doi: 10.1080/17524032.2017.1394892.

Ralf Barkemeyer, Frank Figge, Andreas Hoepner, Diane Holt, Johannes Marcelus Kraak, and Pei-Shan Yu. Media coverage of climate change: An international comparison. *Environment and Planning C: Politics and Space*, 35(6):1029–1054, 2017.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *ICWSM*, 2020.

Melgris José Becerra, Marcia Aparecida Pimentel, Everaldo Barreiros De Souza, and Gabriel Ibrahin Tovar. Geospatiality of climate change perceptions on coastal regions: A systematic bibliometric analysis. *Geography and Sustainability*, 1(3):209–219, September 2020. doi: 10.1016/j.geosus.2020.09.002. URL `https://doi.org/10.1016/j.geosus.2020.09.002`.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. *Proceedings of the Annual ACM Conference on Computational Learning Theory*, 10 2000. doi: 10.1145/279943.279962.

M. Boykoff and Jules Boykoff. Balance as bias: global warming and the us prestige press. *Global Environmental Change-human and Policy Dimensions*, 14:125–136, 2004.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. The MIT Press, 2006. ISBN 9780262033589. URL `http://dblp.uni-trier.de/db/books/collections/CSZ2006.html`.

Emily M. Cody, A. J. Reagan, Lewis Mitchell, P. Dodds, and C. Danforth. Climate change sentiment on twitter: An unsolicited public opinion poll. *PLoS ONE*, 10, 2015.

Diyang Cui, Shunlin Liang, and Dongdong Wang. Observed and projected changes in global climate zones based on köppen climate classification. *Wiley Interdisciplinary Reviews: Climate Change*, 12 (3):e701, 2021.

Biraj Dahal, S. Kumar, and Zhenlong Li. Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining*, 9:1–20, 2019.

Amitava Das and Björn Gambäck. Identifying languages at the word level in code-mixed Indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387, Goa, India, December 2014. NLP Association of India. URL https://aclanthology.org/W14-5152.

Riddhiman Dasgupta, Balaji Ganesan, Aswin Kannan, Berthold Reinwald, and Arun Kumar. Fine grained classification of personal data entities. *arXiv preprint arXiv:1811.09368*, 2018.

J. Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613 – 619, 1973.

Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis R. Bach and David M. Blei, editors, *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189. JMLR.org, 2015. URL http://dblp.uni-trier.de/db/conf/icml/icml2015.html#GaninL15.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *EACL*, 2017.

A. Kirilenko, T. Molodtsova, and S. Stepchenkova. People as sensors: Mass media and local temperature influence climate change discussion on twitter. *Global Environmental Change-human and Policy Dimensions*, 30:92–100, 2014.

Andrei P. Kirilenko and Svetlana O. Stepchenkova. Public microblogging on climate change: One year of twitter worldwide. *Global Environmental Change*, March 2014. ISSN 09593780. doi: 10.1016/j.gloenvcha.2014.02.008. URL http://dx.doi.org/10.1016/j.gloenvcha.2014.02.008.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70, 2022. doi: 10.1109/TKDE.2020.2981314.

S. Linden. Determinants and measurement of climate change risk perception, worry, and concern. *Social Science Research Network*, 2017.

Cedric Lothritz, Kevin Allix, Lisa Veiber, Jacques Klein, and Tegawendé François D Assise Bissyande. Evaluating pretrained transformer-based models on the task of fine-grained named entity recognition.

In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3750–3760, 2020.

M. Loureiro and M. Alló. Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the u.k. and spain. *Energy Policy*, 143:111490, 2020.

Yiwei Luo, D. Card, and Dan Jurafsky. Desmog: Detecting stance in media on global warming. *ArXiv*, abs/2010.15149, 2020.

D. Maynard and Kalina Bontcheva. Understanding climate change tweets: an open source toolkit for social media analysis. In *EnviroInfo/ICT4S*, 2015.

Brendan Moore, Caroline Verfuerth, Angela Minas, Christianne Tipping, Sarah Mander, Irene Lorenzoni, Claire Hoolohan, Andrew Jordan, and Lorraine Whitmarsh. Transformations for climate change mitigation: A systematic review of terminology, concepts, and characteristics. *Wiley Interdisciplinary Reviews: Climate Change*, 12, 11 2021. doi: 10.1002/wcc.738.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 751–760, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587998. doi: 10.1145/1772690.1772767. URL `https://doi.org/10.1145/1772690.1772767`.

W. Pearce, K. Holmberg, I. Hellsten, and B. Nerlich. Climate change on twitter: topics, communities and conversations about the 2013 ipcc report. 2013.

Nicholas Roxburgh, Dabo Guan, Kong Joo Shin, William Rand, Shunsuke Managi, Robin Lovelace, and Jing Meng. Characterising climate change discourse on social media during extreme weather events. *Global Environmental Change*, 54:50–60, 2019. ISSN 0959-3780. doi: https://doi.org/10.1016/j.gloenvcha.2018.11.004. URL `https://www.sciencedirect.com/science/article/pii/S0959378018305776`.

Kamol Chandra Roy, Samiul Hasan, Arif Mohaimin Sadri, and Manuel Cebrian. Understanding the efficiency of social media based crisis communication during hurricane sandy. *International Journal of Information Management*, 52:102060, June 2020. doi: 10.1016/j.ijinfomgt.2019.102060. URL `https://doi.org/10.1016/j.ijinfomgt.2019.102060`.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

T Tulasi Sasidhar, Premjith B, and Soman K P. Emotion detection in hinglish(hindi+english) code-mixed social media text. *Procedia Computer Science*, 171:1346–1352, 2020a. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2020.04.144. URL `https://www.sciencedirect.com/`

science/article/pii/S1877050920311236. Third International Conference on Computing and Network Communications (CoCoNet'19).

T Tulasi Sasidhar, Premjith B, and Soman K P. Emotion detection in hinglish(hindi+english) code-mixed social media text. *Procedia Computer Science*, 171:1346–1352, 2020b. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2020.04.144. URL https://www.sciencedirect.com/science/article/pii/S1877050920311236. Third International Conference on Computing and Network Communications (CoCoNet'19).

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus.

Parinaz Sobhani, Saif M. Mohammad, and Svetlana Kiritchenko. Detecting stance in tweets and analyzing its interaction with sentiment. In **SEMEVAL*, 2016.

Varsha Suresh and Desmond C Ong. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. *arXiv preprint arXiv:2109.05427*, 2021.

Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. An english-hindi code-mixed corpus: Stance annotation and baseline system. *arXiv preprint arXiv:1805.11868*, 2018.

Pradip Swarnakar and Ashutosh Modi. Nlp for climate policy: Creating a knowledge platform for holistic and effective climate action. *ArXiv*, abs/2105.05621, 2021.

Hywel TP Williams, James R McMurray, Tim Kurz, and F Hugo Lambert. Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global environmental change*, 32:126–138, 2015.